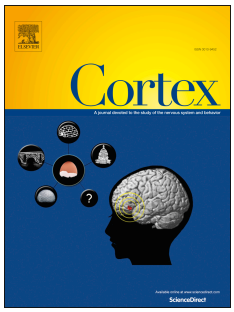


Intermodulation frequencies reveal common neural assemblies integrating facial and vocal fearful expressions

Francesca M. Barbero, Siddharth Talwar, Roberta P. Calce, Bruno Rossion, Olivier Collignon

PII: S0010-9452(24)00347-2

DOI: https://doi.org/10.1016/j.cortex.2024.12.008

Reference: CORTEX 4069

To appear in: *Cortex*

Received Date: 13 August 2024

Revised Date: 18 November 2024

Accepted Date: 2 December 2024

Please cite this article as: Barbero FM, Talwar S, Calce RP, Rossion B, Collignon O, Intermodulation frequencies reveal common neural assemblies integrating facial and vocal fearful expressions, *CORTEX*, https://doi.org/10.1016/j.cortex.2024.12.008.

Intermodulation frequencies reveal common neural assemblies integrating facial and vocal fearful expressions

Author names and affiliations:

Francesca M. Barbero [a], Siddharth Talwar [a], Roberta P. Calce [a], Bruno Rossion [b,c], Olivier Collignon [a,d]

[a] Institute of Research in Psychology (IPSY) & Institute of Neuroscience (IoNS), Louvain Bionics Center, University of Louvain (UCLouvain), 1348 Louvain-la-Neuve, Belgium

[b] Université de Lorraine, CNRS, IMoPA UMR 7365, F-54000 Nancy, France

[c] Université de Lorraine, CHRU-Nancy, Service de Neurologie, F-54000, France

[d] School of Health Sciences, HES-SO Valais-Wallis, The Sense Innovation and Research Center, 1007 Lausanne & 1950 Sion, Switzerland

Corresponding authors:

Correspondence should be addressed to Francesca M. Barbero and Olivier Collignon, Institute of Research in Psychology (IPSY) & Institute of Neuroscience (IoNS), Louvain Bionics Center, University of Louvain (UCLouvain), 1348 Louvain-la-Neuve, Belgium. E-mail addresses: francesca.barbero@uclouvain.be and olivier.collignon@uclouvain.be

Declaration of interest: The authors declare no competing interests.

Intermodulation frequencies reveal common neural assemblies integrating facial and vocal fearful expressions

Author names and affiliations:

Francesca M. Barbero [a], Siddharth Talwar [a], Roberta P. Calce [a], Bruno Rossion [b,c], Olivier Collignon [a,d]

[a] Institute of Research in Psychology (IPSY) & Institute of Neuroscience (IoNS), Louvain Bionics Center, University of Louvain (UCLouvain), 1348 Louvain-la-Neuve, Belgium

[b] Université de Lorraine, CNRS, IMoPA UMR 7365, F-54000 Nancy, France

[c] Université de Lorraine, CHRU-Nancy, Service de Neurologie, F-54000, France

[d] School of Health Sciences, HES-SO Valais-Wallis, The Sense Innovation and Research Center, 1007 Lausanne & 1950 Sion, Switzerland

Corresponding authors:

Correspondence should be addressed to Francesca M. Barbero and Olivier Collignon, Institute of Research in Psychology (IPSY) & Institute of Neuroscience (IoNS), Louvain Bionics Center, University of Louvain (UCLouvain), 1348 Louvain-la-Neuve, Belgium. E-mail addresses: francesca.barbero@uclouvain.be and olivier.collignon@uclouvain.be

Declaration of interest: The authors declare no competing interests.

Abstract

Effective social communication depends on the integration of emotional expressions coming from the face and the voice. Although there are consistent reports on how seeing and hearing emotion expressions can be automatically integrated, direct signatures of multisensory integration in the human brain remain elusive. Here we implemented a multi-input electroencephalographic (EEG) frequency tagging paradigm to investigate neural populations integrating facial and vocal fearful expressions. High-density EEG was acquired in participants attending to dynamic fearful facial and vocal expressions tagged at different frequencies (f_{vis}, f_{aud}). Beyond EEG activity at the specific unimodal facial and vocal emotion presentation frequencies, activity at intermodulation frequencies (IM) arising at the sums and differences of the harmonics of the stimulation frequencies ($mf_{vis} \pm nf_{aud}$) were observed, suggesting non-linear integration of the visual and auditory emotion information into a unified representation. These IM provide evidence that common neural populations integrate signal from the two sensory streams. Importantly, IMs were absent in a control condition with mismatched facial and vocal emotion expressions. Our results provide direct evidence from non-invasive recordings in humans for common neural populations that integrate fearful facial and vocal emotional expressions.

Keywords

Multisensory integration; Emotion; Face-voice integration; Intermodulation frequency; Electroencephalography

1. Introduction

Our ability to efficiently recognize emotional expressions is arguably one of the most important perceptual skills for successful social interactions in humans and other animals (Darwin, 1872). Even if most research focused on either facial or vocal emotional expressions, emotion expressions perception and production are intrinsically multisensory (de Gelder & Vroomen, 2000): we are more accurate and faster in discriminating emotions when congruent auditory and visual cues are available (i.e., depicting the same emotion expression; Collignon et al., 2008; Ethofer, Pourtois, et al., 2006; Falagiarda & Collignon, 2019) and we tend to rely on information from both modalities to make a decision on the perceived emotion even when explicitly asked to ignore one sensory channel (Campanella & Belin, 2007; Collignon et al., 2008; de Gelder & Vroomen, 2000; Dolan et al., 2001; Ethofer, Anders, et al., 2006; Klasen et al., 2012; Massaro & Egan, 1996).

Although concurrent congruent information from the face and the voice improves perceptual discrimination, whether and if so how the brain creates a unified representation of affective information delivered from the face and the voice remains unknown. Electrophysiological studies in nonhuman primates have shown that single neurons in primary and associative cortices integrate facial and vocal signals from conspecifics (Barraclough*, Xiao* et al., 2005; Ghazanfar et al., 2005; Ghazanfar & Schroeder, 2006; Khandhadia et al., 2021; Perrodin et al., 2014, 2015; Romanski, 2007, 2012; Stein & Stanford, 2008; Sugihara et al., 2006). In humans, functional neuroimaging studies have investigated the neural correlates of emotion integration by probing how unimodal and multimodal emotion processing differ in specific brain regions (Davies-Thompson et al., 2019; Dolan et al., 2001; Ethofer, Anders, et al., 2006; Ethofer, Pourtois, et al., 2006;

Klasen et al., 2012; Peelen et al., 2010; Pourtois et al., 2005; Watson et al., 2014) and at specific time-points (de Gelder et al., 1999, 2002; Föcker & Röder, 2019; Ho et al., 2015; Jessen & Kotz, 2011; Kokinous et al., 2015; Pourtois et al., 2000). However, revealing a direct signature of multisensory integration remains challenging**,** particularly in the human brain. On the one hand, functional Magnetic Resonance Imaging (fMRI) has identified brain regions that respond differently to multimodal than unimodal emotion expressions, but cannot determine whether there are common neural populations that respond to both facial and vocal emotion information. This is because neural activity recorded at each voxel comes from millions of neurons (Logothetis, 2008), making it impossible to disentangle between the possibilities that higher responses to audio-visual emotion rather than to auditory or visual emotion alone are driven either by: a) populations of bimodal neurons or b) populations of unimodal neurons that respond to emotional faces or voices and coexist in the same voxel (Stanford & Stein, 2007). On the other hand, magnetic resonance imaging and electrophysiological studies relying on the rejection of the additive null hypothesis (e.g., audiovisual ≠ audio + visual) to demonstrate multisensory integration do not consider that the additive equation is unbalanced, as all processes that are not strictly related to the signal of interest and that are part of the neural response (e.g., salience, arousal, noise) would be counted once for the bimodal stimuli but only once for the unimodal stimuli (Giani et al., 2012; Gondan & Röder, 2006; Stevenson et al., 2014).

To overcome these limitations, we propose a multi-input frequency-tagging technique as a non-invasive approach to investigate whether there are neural populations in the human brain that integrate facial and vocal emotion expressions depicting fear. We recorded EEG while participants attended audiovisual clips where dynamic fearful facial

and vocal expressions were presented concurrently at different frequencies (f_1 and f_2), with the rationale that if there are neural populations that integrate fearful information from the face and the voice, we should observe activity at intermodulation frequencies ($mf_1 \pm nf_2$) (Baldauf & Desimone, 2014; Boremanse et al., 2013, 2014; Giani et al., 2012; Gordon et al., 2019; Norcia et al., 2015; Regan et al., 1995).

2. Material and methods

2.1. Participants

We acquired electrophysiological (EEG) recordings from twenty participants that were naïve about the goal of the study. Sample size was chosen in accordance with previous studies using frequency-tagging to investigate emotion perception from the face or the voice (e.g., Dzhelyova et al., 2016: n = 18; Leleu et al., 2018: n = 18; Matt et al., 2021: n = 17; Talwar et al., 2023: n = 24). During data preprocessing, one participant was excluded due to the presence of massive artefacts in the EEG trace, leading to a final sample of nineteen participants (9 females, mean age = 25.32 years, SD = 5.07 years). All participants reported to have normal or corrected-to-normal vision, normal hearing and no history of neurological disorders. The experimental procedure was approved by the local ethical committee of the University of Louvain (project 2016-25) and conducted in accordance with the Declaration of Helsinki. All participants received financial compensation for participating in the experiment.

2.2. Stimuli

Stimuli were selected from a collection of audiovisual clips in which actors portrayed the prototypical expressions of different emotions (anger, disgust, fear, happiness, sadness, surprise, neutral) and pain. During the recording session, professional actors were

instructed to produce simultaneous facial and vocal emotion expressions. The facial expressions were "prototypical" insofar as they possessed the key features (identified using the Facial Action Coding System: FACS) identified by Ekman and Friesen (1976) as being representative of everyday facial expressions (actors were trained to produce corresponding FACS features linked to each emotion). The vocal expressions were short, non-verbal emotional interjections using the French vowel 'ah' (Belin et al., 2008). A selection of the audiovisual clips has been previously presented in other experiments multimodally (audio-visual; Falagiarda & Collignon, 2019) or unimodally (visual-only; Simon et al., 2006), and a subset of the auditory-only vocalization is available in the Montreal Affective Voices dataset (Belin et al., 2008). Based on the results of a pilot study, we selected an actress whose emotion expressions were efficiently discriminated by the vast majority of the participants tested (high accuracy score and no biases in the incorrect answers, i.e., no specific emotion systematically confused with others among the considered expressions), selecting an audiovisual clip depicting fear and one depicting disgust. We selected fear and disgust as, from an evolutionary perspective, the efficient processing of threat related signals has been suggested to be beneficial for survival (T. Stein et al., 2014). Fear and disgust do however signal different kinds of environmental threats (e.g., impending attack and potential contamination or poisoning) and prompt distinct responses (e.g., fight or flight response versus further inspection or avoidance; Anderson et al., 2003). Crucially, notwithstanding the fact that both fearful and disgust emotion expressions have a negative valence, they can be clearly discriminated from one another (Collignon et al., 2008).

The original clips had a frame rate of 29.97 fps and a duration of 1 second. We extracted the auditory and the visual information from the clips in order to be able to present them

independently. First, we cropped the frames of the fear clip using iMovie 10.1.14 running on macOS Mojave 10.14.6 to remove part of the neutral background and center the frame around the face of the actress. Then, single frames were extracted using custom built script running on Matlab R2016b (MathWorks). The resulting frames had a dimension of 1088 x 1072 pixels. The audio files were extracted from the fear and the disgust clips using QuickTime Player 10.5. During the procedure, a sampling rate of 32 kHz instead of 30 kHz was erroneously assigned to the auditory fear and disgust vocalizations, which resulted in vocalizations sounding slightly higher in pitch, but still ecologically valid and plausible. The extracted audio files were then edited using Audacity 2.1.2 (https://audacityteam.org/): each audio file was first cropped to remove the silence at the beginning of the vocalization and then transformed from stereo to mono. Resulting files were saved as wav files with a sampling rate of 32 kHz and a duration of 808.9 ms for auditory disgust and 961.3 ms for auditory fear.

2.3. Choice of stimulation frequencies

Since we used a multi-input frequency tagging approach to investigate audio-visual integration of emotion expressions, the auditory and the visual information were presented, or 'tagged', at different frequencies. The choice of the stimulation frequencies was motivated by different factors. First, we wanted the auditory stimulation cycle to be longer in time than the visual one ($t_{aud} > t_{vis}$ and hence $f_{aud} < f_{vis}$). This is because emotion recognition is more accurate and faster in vision than in audition (Bänziger et al., 2009; Falagiarda & Collignon, 2019), a factor that might explain why we tend to rely more on vision than audition to efficiently discriminate emotions (Collignon et al., 2008; Klinge et al., 2010; Watson et al., 2014). Therefore, we presented the auditory emotion information for a longer time in an effort to make the information available through faces

and voices equally informative (i.e. suppression of the advantage of vision by presenting visual information for a shorter time) (Falagiarda & Collignon, 2019) as, according to the principle of "inverse effectiveness" (B. E. Stein & Meredith, 1993), multisensory integration is more effective when the single modalities are less salient. Second, inspired by studies using intermodulation frequencies to investigate high-level intramodal visual integration processes (Boremanse et al., 2013, 2014), we selected two temporally close stimulation frequencies. Lastly, as common practice in frequency tagging experiments, we selected stimulation frequencies that are not too low in order to avoid falling into a region of high electrophysiological noise in the EEG spectrum (Luck, 2014). Considering all these factors, we presented the auditory emotion information for 500 ms (full cycle 1000 ms: stimulus played forward and then backward) and the visual emotion for 333 ms (full cycle 666 ms: unfolding of emotion forward and then backward). This resulted in a multi-input audio-visual stimulation at 1 Hz (auditory) and 1.5 Hz (visual) that was resetting (i.e., restarting in sync) every 2 seconds (2 full auditory cycles and 3 full visual cycles).

2.4. Experimental design

The experiment consisted of the concurrent presentation of auditory and visual emotion information which could be either congruent (same emotion) or not (different emotion). In the congruent condition participants were simultaneously presented with visual fear and auditory fear, while in the incongruent condition visual fear was paired with auditory disgust. It is important to note that: (i) in the congruent condition, the facial and vocal expression come from the same original bimodal recording, and that (ii) the pairing of visual fear with auditory disgust was just one of the possible manipulations to obtain incongruency between facial and vocal expressions. As we used a multi-input frequency

tagging approach to investigate audio-visual integration of fearful expressions, the auditory and the visual information were presented, or 'tagged', at different frequencies: 1 and 1.5 Hz respectively. To do so, depending on the trial condition, the auditory file depicting fear or disgust was cropped to include the first 500 ms and equalised in overall energy to a value of 0.05 online. A 10 ms ramp was applied to the stimulus to avoid clicking and avoid abrupt onset-offset of the sound. The auditory stimulation consisted of presenting in succession the auditory file forward and backward for the whole duration of the trial. To tag the unfolding of visual emotion expression at 1.5 Hz, we presented the first 11 (out of 30) frames of the audio-visual clip for fear at a rate of 30 Hz to be as close as possible to the frame rate of the original videos while respecting the constraints of the refresh rate of the monitor (120 Hz). The frames were presented forward and backward for the whole duration of the trial, and to facilitate the perception of a smooth unfolding of the emotion expression (forward, backward and forward again), the first and the last frame were presented only once in a presentation cycle (i.e., for each presentation cycles we presented frames in the order: 1 2 3 4 5 6 7 8 9 10 11 10 9 8 7 6 5 4 3 2). Therefore, the time to reach the apex of fear facial expression from the neutral face was 333 ms (full cycle: 666 ms, frequency = 1/time full cycle = 1.5 Hz). The frames were displayed in the middle of the screen on a black background. With participants seating at 80 cm distance from the screen, the frames subtended 7° 15' 0.55" x 7° 9' 0.16" of visual angle. Each trial had a duration of 64 seconds, including 2 s of fade-in and 2 s of fade-out for the gradual increase/decrease of luminance of the visual stimulation. Triggers corresponding to the onset of auditory stimulation and to the beginning of each visual stimulation cycle were sent from the stimulation computer to the EEG amplifier. A schematic representation of the experimental design is shown in figure 1. The experiment consisted of the presentation

of 20 trials (each of 64 s): 10 for each condition. In the congruent condition, visual fear and auditory fear were presented concurrently. In the incongruent condition, visual fear was presented simultaneously with auditory disgust to disrupt potential audio-visual integration processes which might occur in the presence of congruent (i.e., same emotion) audio-visual information. The two conditions were presented in a pseudo-randomised order. Demos for the congruent and the incongruent conditions are available here: https://osf.io/hqvrc/?view_only=a9fd413e05c34ba5aab316a67c661e8a. To ensure attention to the auditory and the visual modality, participants were instructed to perform an orthogonal bi-modal task consisting in detecting transient decrease in luminance or volume in the visual or auditory stimulation respectively (four occurrences in each trial) through button press. The auditory attentional targets were obtained by decreasing of a factor of 3 the RMS value of two individual sounds across the trial. The visual attentional targets were introduced by decreasing the luminance of the visual stimulation to half for 500 ms (to mirror the duration of the auditory attentional target) twice in a trial. The experiment was implemented in Matlab R2016b using the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). Visual stimulation was presented on a Benq XL2420T monitor with a resolution of 1920 x 1080 pixels and a refresh rate of 120 Hz. The auditory stimulation was presented through Logitech Z200 speakers, with the sound intensity kept at a comfortable level of approximately 70 dB. Participants performed the experiment in a quiet, dimly lit room seated at 80 cm from the monitor and at 85 cm from the speakers while we recorded brain activity using a 128-channels BioSemi Active Two EEG system (https://www.biosemi.com/products.htm). Recording sites included standard 10-20 system locations as well as intermediate positions (channel coordinates can be found at https://www.biosemi.com/headcap.htm).

Electrodes offset relative to the common mode sense (CMS) and driven right leg (DRL) electrode loop was kept below ± 30 mV. The EEG was digitized at a sample rate of 512 Hz.

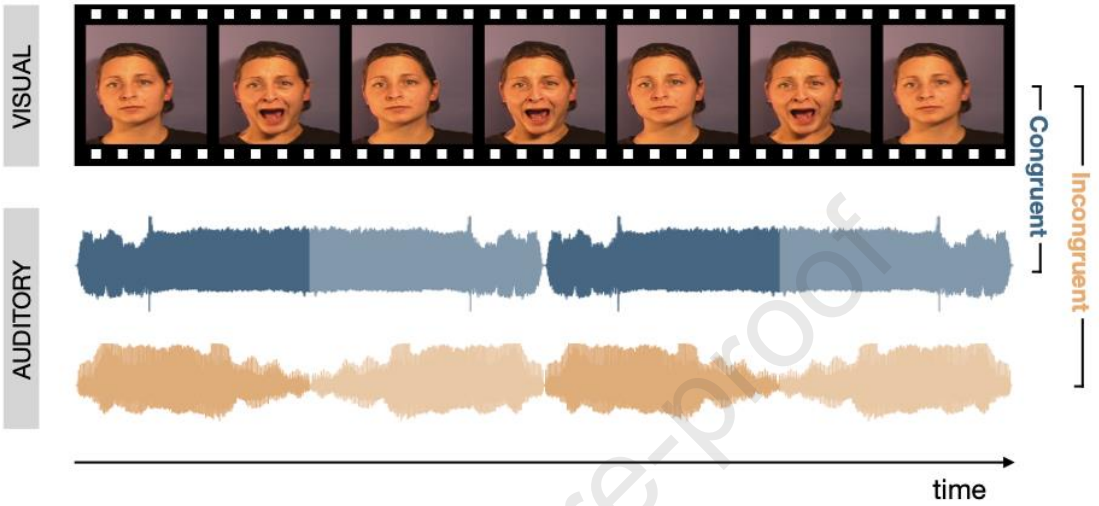


Figure 1. Experimental design. Dynamic fearful facial expressions were presented simultaneously with an affective vocalization expressing either fear (congruent condition) or disgust (incongruent condition). To present the visual and auditory affective expressions periodically, they were presented forward and then backwards (in lighter color for the auditory stimulation) in a loop for the duration of the sequences. The schematic representation corresponds to 2-seconds excerpts of a trial.

2.5.EEG data preprocessing

Data analysis was performed using Letswave 6 (https://github.com/NOCIONS/letswave6) running on Matlab R2016b, Matlab R2022a (MathWorks) and RStudio (https://www.rstudio.com) (Rstudio, 2020) following procedures validated in many EEG frequency-tagging studies (e.g., Barbero et al., 2021; Calce et al., 2024; Retter & Rossion, 2016; Talwar et al., 2023; Volfart et al., 2021). The raw EEG data were bandpass filtered between 0.1 and 100 Hz with a fourth order Butterworth bandpass filter to remove slow drifts and very high frequencies and then down sampled to 256 Hz to facilitate data handling and storage. The continuous data were

then segmented from -2 s to +66 s with respect to the onset of the audio-visual stimulation before performing artefact rejection. We used independent component analysis (ICA) using the RUNICA algorithm as implemented in EEGLAB using the square method to identify and remove artefactual components due to blinks and eye movements, removing no more than 2 components in each participant (1 component removed in 19 participants, 2 components in 1 participant). After visual inspection of the data, residual noisy electrodes (less than 3 out of the 128 channels per participant, corresponding to less than 3% of the electrodes: 1 electrode in 5 participants, 3 electrodes in 2 participants and no electrodes in the remaining 12 participants) were linearly interpolated and 1 noisy epoch was removed. At this stage, one participant was excluded from further analysis due to the massive presence of artefacts on the EEG trace. The data of the remaining 19 participants were re-referenced to the common average and re-segmented from +2 s to +62 s relative to stimulation onset (thus excluding fade-in and fade-out) separately for each participant and condition. Resulting epochs contained a number of bins that corresponded to integer multiples of the stimulation (f_{aud}, f_{vis}) and intermodulation frequencies components ($mf_{aud} \pm nf_{vis}$) to avoid data overspill into neighbouring frequencies during the Fourier transformation. To obtain a higher resolution in the frequency domain, we concatenated epochs 2-by-2 separately for each participant and condition. Therefore, from the initial 10 x 60 seconds epochs per condition, we obtained 5 x 120 seconds epochs (frequency resolution = 1 / epoch duration; i.e., 1/120 s = 0.00833 Hz). For each condition and participant separately, epochs were averaged in the time domain to attenuate EEG activity not in phase with the stimulation. Averaged epochs were submitted to a Fast Fourier Transform (FFT) leading to amplitude spectra for each electrode, participant and condition.

2.6. Frequency domain and statistical analysis

2.6.1. Unisensory responses

To define the number of significant harmonics in each modality and condition (i.e., for auditory fear and disgust responses at f_{aud}, $2f_{aud}$, …), the amplitude spectra were grand-averaged across participants for each condition separately. We then pooled together all 128 electrodes and we calculated the z-score at each frequency bin of this averaged channel. For each bin, the z-score was calculated as the difference between the amplitude at the bin of interest and the averaged amplitude of 22 surrounding bins (11 on each side), excluding the immediately adjacent bins, the local minimum and the local maximum to avoid potential remaining spectral leakage and to avoid projecting signal into neighbouring bins containing noise, divided by the standard deviation of the same selected bins, as common practice in frequency-tagging studies analyses pipelines (Retter & Rossion, 2016; Volfart et al., 2021). Consecutive harmonics with a z-score value higher than 1.64 (i.e., $p < 0.05$, 1-tailed, signal > noise) were considered significant. To quantify unisensory responses, we summed responses at the auditory and visual frequencies and significant harmonics, excluding the harmonics that were shared among auditory and visual stimulation. To do so, we first segmented individual FFT epochs into chunks centred on the frequencies of interest for the auditory and visual responses in congruent and incongruent conditions separately. Each chunk contained 25 bins: the frequency of interest and 12 bins on each side, with a total length of 0.2 Hz (0.1 Hz on each side). The chunks were summed separately for each electrode, participant, modality and condition and then grand-averaged across participants. Three different operations were computed on grand-averaged chunked responses: baseline correction of the amplitude, z-score computation and estimation of the signal-to-noise ratio (SNR) of the responses. For the

baseline correction, we subtracted from the amplitude at the bin of interest the average amplitude of the 22 surrounding bins excluding the adjacent bins and the local minimum and maximum ($x - bl$; baseline calculated in the same way as for the z-scores as reported above). Z-scores were calculated using the same operation described in the procedure to select significant harmonics ($zscore = \frac{x-bl}{std(bl)}$). Finally, the signal-to-noise ratio (SNR) of the responses was obtained dividing the amplitude of the bin of interest by the average amplitude of the 22 surrounding bins (excluding the two adjacent bins, one minimum and one maximum, $SNR = x/bl$). Significant electrodes were assessed from z-score values of the grand-averaged responses to emotional faces and voices, correcting for multiple comparisons using the False Discovery Rate (FDR) method (Benjamini & Hochberg, 1995). To investigate crossmodal effects on unisensory responses, we calculated the difference between the grand-averaged amplitudes (FFT) to emotional faces in the congruent and incongruent conditions, for which the visual stimulation was identical. This differential response was then submitted to z-score calculation and a FDR correction was implemented to control for multiple comparisons.

2.6.2. Crossmodal responses

To investigate whether we could isolate significant intermodulation frequency responses ($mf_{aud} \pm nf_{vis}$) reflecting audio-visual emotion integration, we first averaged amplitude spectra across participants and electrodes for the congruent and incongruent condition separately. Then, we calculated the difference between the amplitude of the obtained signals for the two conditions, computed the z-score at each frequency bin as described above and considered its absolute value (not to introduce a bias toward which condition would have driven a significant response). Subtraction across conditions was done to isolate responses that would truly reflect category-selective integration of bimodal

emotions as it has been reported that some brain areas respond similarly to emotionally congruent and incongruent bimodal stimuli (Klasen et al., 2012). Intermodulation frequency activity was considered as significant if their z-score's absolute value was higher than 1.64 ($p < 0.05$). Intermodulation frequencies overlapping with harmonics of the unimodal frequencies of stimulation (e.g., 2 Hz = 2* f_{aud} = 2*f_{vis} - f_{aud} = 2*1.5 Hz - 1 Hz) were not included in the analysis as it is impossible to disentangle responses driven by the unimodal stimulation and integration responses at these frequencies. Once identified the significant harmonics, to quantify the crossmodal response we followed the same procedure implemented for the unisensory responses. Individual FFT epochs were segmented into chunks centred on the selected frequencies for the congruent and incongruent conditions separately using the same parameters (25 bins: frequency of interest and 12 frequency bins on each side). Resulting epochs were summed separately for each electrode, participant and condition. Baseline correction of the amplitude, z-score computation and SNR estimation were performed on individual and grand-averaged data. For the two conditions and their differential response (congruent-incongruent), significant electrodes were defined from the z-scores of the grand-averaged responses using the FDR method to correct for multiple comparisons.

An additional region-of-interest (ROI) analysis was performed to compare individual crossmodal responses across the congruent and incongruent conditions. We a priori defined two regions-of-interest based on previous EEG frequency tagging studies on visual (Leleu et al., 2018) and auditory (Talwar et al., 2023) emotion processing. The visually-defined ROI (vROI) comprised 19 contiguous occipito-parietal electrodes (P9, P7, PO7, PO9, PO11, I1, POI1, O1, Oz, Oiz, Iz, I2, POI2, O2, PO8, PO10, PO12, P8, P10; figure 3C). The auditory-defined ROI (aROI) consisted of 20 contiguous fronto-

central electrodes (Cz, C1h, FCC1h, FCC2h, C2h, C1, FCC1, FCz, FCC2, C2, FC3h, FFC1, FFCz, FFC2, FC4h, FFC3h, F1, Fz, F2, FFC4h; figure 3C). For each participant, electrode, and condition, we segmented FFT epochs into chunks centered on the intermodulation frequencies identified as significant in the previous analysis. The segmented epochs were comprised of 25 bins: the frequency of interest and 12 frequency bins on each side. For each participant, condition, and region-of-interest, we averaged the resulting epochs across the selected electrodes. The responses were then submitted to SNR estimation by dividing the amplitude response at the bin of interest by the average amplitude of the 22 surrounding bins (excluding the two adjacent bins, one minimum and one maximum, $SNR = x/bl$). Paired t-tests were used to investigate an effect of condition on individual responses. We corrected for multiple comparisons using the False Discovery Rate (FDR) method (Benjamini & Hochberg, 1995); threshold for statistical significance was set to 0.05.

3. Results

3.1. Responses to emotional voices

Responses were expected at the auditory stimulation frequency and harmonics (Retter et al., 2021), reflecting the processing of affective vocalizations and general auditory processes. Excluding responses at harmonics that were shared among the auditory and visual stimulation (e.g., $3f_{aud} = 2f_{vis} = 3$ Hz), we were able to observe significant consecutive harmonics up to 30 Hz depending on the condition. However, since responses at harmonics higher than 10 Hz were associated with low SNR, we focused our subsequent analyses on responses up to 10 Hz (i.e., 85.85 % of the baseline-corrected amplitude of the responses until 30 Hz at auditory, visual and intermodulation frequencies

averaged across all electrodes, participants, and conditions). The SNR spectra obtained from the average of all electrodes of grand-averaged data is presented in figure 2B for visualization purposes.

To quantify responses to auditory fear and disgust, we summed responses at 1, 2, 4, 5, 7, 8 and 10 Hz for each participant and condition separately. Individual responses were then grand-averaged and three different operations were performed: baseline subtraction, z-scores calculation and SNR estimation. Responses peaked over central, temporo-occipital electrodes with the largest amplitudes found over posterior-temporal electrodes (maximal baseline corrected amplitude at P10: 1.616867 µV for congruent and at TP8: 0.996568 µV for incongruent; figure 2A). Auditory responses were associated with high SNR (congruent: M = 2.549968 ± 0.710588, incongruent: M = 1.972000 ± 0.343552, mean ± standard deviation across all electrodes) and all electrodes showed a significant response in both conditions (FDR corrected; congruent: z-scores' range: 8.706837 - 86.017570, corrected $p \leq 1.5624 \times 10^{-18}$; incongruent: z-score's range: 6.775906 - 54.222603, corrected $p \leq 6.1815 \times 10^{-12}$).

3.2. Responses to emotional faces

Significant responses were expected for facial emotion expressions at the stimulation frequency and related harmonics. Without including harmonics which were also multiples of the auditory stimulation frequency, we were able to isolate consecutive significant harmonics up to 19.5 Hz depending on the condition. As for auditory responses, we focused our analysis on harmonics up to 10 Hz: 1.5, 4.5 and 7.5 Hz (clear peaks at these frequencies can be observed in figure 2B). The sum of these 3 harmonics yielded to the characterization of visual responses with medial and right occipital topographies (figure 2A), reflecting brain responses to fearful facial expressions as well as more general low-

level visual processes. The largest amplitudes were observed over medial and right occipital electrodes (maximum value for baseline corrected amplitude at Oz: 0.560739 µV for congruent and at PO10: 0.576443 µV for incongruent). Visual responses had a high SNR (congruent: M = 2.157388 ± 0.433499, incongruent: M = 2.061532 ± 0.400881 across all electrodes), and even after correcting for multiple comparisons with the FDR method, all electrodes showed a significant response (congruent: z-scores' range: 4.976463 - 46.975430, FDR-corrected $p \leq 3.2378 \times 10^{-7}$; incongruent: z-scores' range: 6.177848 - 45.166496, FDR-corrected $p \leq 3.2490 \times 10^{-10}$). Since the influence of concomitant redundant affective information has been reported to enhance the amplitude of unisensory responses (Pourtois et al., 2000), we investigated whether there was an effect of congruency of the auditory information on the visual response. We computed the difference between the grand-averaged FFT spectra in the congruent and incongruent conditions, and then calculated the baseline corrected amplitudes and the z-scores of this differential response. The difference highlighted higher responses for the congruent condition over right/centro-frontal and left-medial occipital electrodes, showing cross-modal influences on unisensory responses, with 38 electrodes (i.e. 29.6875%) remaining significant after FDR correction (1-tailed t-test, significant electrodes z-scores' range: 2.1813 - 6.5775, significant electrodes FDR-corrected p-values range: 3.0620×10^{-9} - 0.0491; significant electrodes: PPO3, PPO5, PO7, PO9, PO11, I1, POI1, O1, POO5, PO3h, PPOz, POz, POOz, Oz, Iz, POO6, CCP4, CCP2, C2, C4, C6h, C6, FC6, FC6h, FC4, FC4h, FCC2, FFC4h, FFC4, AFF6h, FFC2, F2, AFF2, AFF4h, Fpz, Afz, AFFz, Fz; figure 2C).

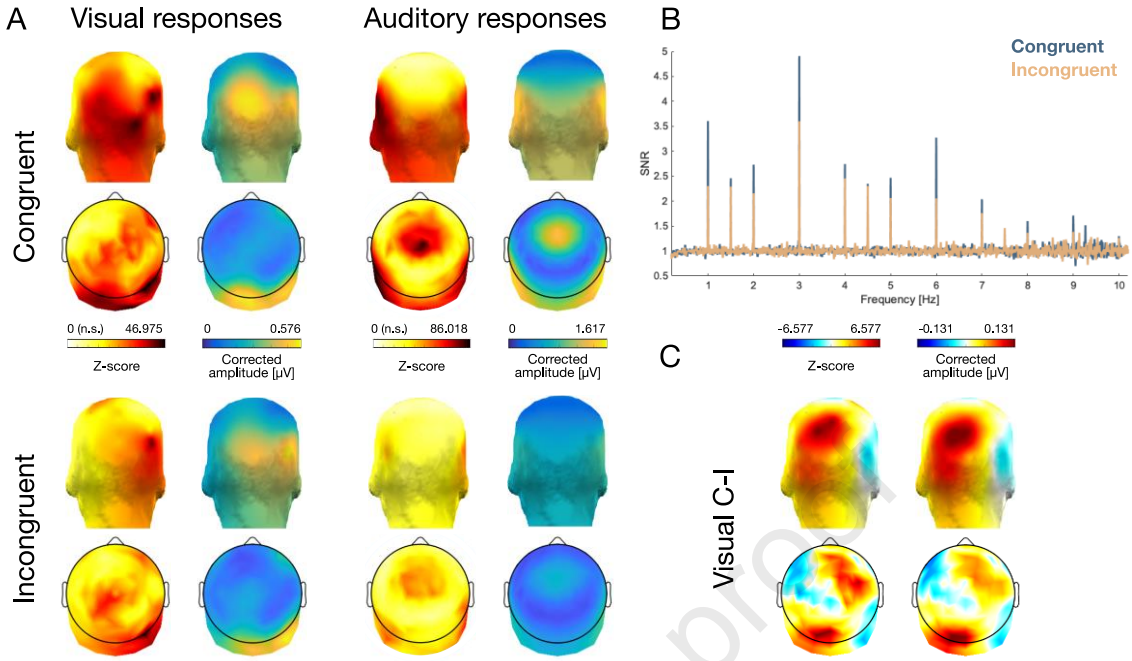


Figure 2. Responses to emotional faces and voices. A) Topographical responses to visual (left) and auditory (right) emotion expressions to the congruent (top) and incongruent (bottom) conditions. On the left of each panel, the 3D and 2D maps represent the distribution of the z-scores of the responses. On the right of each panel, topographies correspond to the baseline corrected amplitudes of the responses. B) SNR spectra of the average across all participants and electrodes show clear responses to facial and vocal affective expressions at the frequencies of stimulation and harmonics for the congruent (blue) and incongruent (orange) conditions. C) Difference maps obtained by subtracting grand-averaged EEG responses to visual stimulation in the congruent and incongruent conditions. 38 out of 128 electrodes showed a significantly higher response when congruent fearful facial and vocal expressions were presented simultaneously.

3.3. Intermodulation frequency responses to multisensory fearful expressions

If there are neural populations that integrate fearful expressions from faces and voices (and whose responses are measurable with EEG; Gordon et al., 2019), we should observe responses at intermodulation frequencies (e.g., $mf_{vis} \pm nf_{aud}$). Such intermodulation

frequencies are elicited if there is a population of neurons that receive inputs from the two stimuli, and can be considered as an index of nonlinear integration (Baldauf & Desimone, 2014; Boremanse et al., 2013, 2014; Giani et al., 2012; Gordon et al., 2019; Norcia et al., 2015; Regan et al., 1995).

For this analysis, we excluded frequency bins shared between intermodulation frequency components and harmonics of the auditory and/or visual stimulation (e.g., $2f_{vis} - f_{aud} = 2f_{aud} = 2$ Hz), as it is not possible to disentangle between responses elicited by the unimodal stimulation and integration responses at these frequencies. It has to be noted that while some activity at intermodulation frequency components reflecting integration could be present at harmonics considered for the analysis of unisensory responses (as for 2 Hz, for instance), these frequency bins were nonetheless included in the analysis of unimodal auditory and visual responses since: (i) the magnitude of the multisensory responses was expected to be consistently lower than the unisensory ones, therefore only marginally contributing to the quantified auditory and visual responses, (ii) we considered this potential confound not detrimental to our research as unisensory analyses were mostly introduced as a sanity check.

Focusing on the 0-10 Hz frequency range, potentially significant intermodulation frequency responses could be present at: 0.5, 2.5, 3.5, 5.5, 6.5, 8.5 and 9.5 Hz. When probing the differential response between congruent and incongruent stimulation to isolate responses reflecting category specific integration of bimodal emotions expressions (Klasen et al., 2012), we observed significant responses at 0.5, 2.5, 3.5, 6.5 and 9.5 Hz. Responses at the selected intermodulation frequencies were then summed separately for each condition (congruent, incongruent) to quantify multisensory processes. Responses to facial and vocal fearful expressions peaked over bilateral occipital and central

electrodes with the highest response observed over left occipital electrodes (P9: 0.101679 μV, baseline corrected amplitude; figure 3A, top). 18 electrodes (i.e., 14.0625 %) located over central and bilateral (mostly left) occipito-temporal regions remained significant after FDR correction for multiple comparisons (Cz, PPO5, PO7, PO9, POI1, O1, I2, O2, FC4h, FCC2h, FCC1h, FCC1, FC3h, C1h, C3h, C3, P7, P9; significant electrodes z-scores' range: 2.580775 - 4.298480, significant electrodes FDR-corrected p-values range: 0.0011 - 0.0351).

Conversely, no clear multisensory response was evident from the topographical distribution elicited by the simultaneous presentation of incongruent facial and vocal emotion expressions (figure 3A, bottom). In fact, we did not observe any significant electrode after correcting for multiple comparisons and the SNR of the response across the scalp was close to 1 (i.e., signal = noise; M = 0.965419 ± 0.059850 across all electrodes; z-scores' range: -4.6310 - 2.1927). Our results suggest that the multisensory responses elicited from the audio-visual presentation of fearful expressions reflect integration of semantically congruent affective information from the face and the voice as the same response was virtually absent in the control incongruent condition. To further investigate whether there was a difference between the two conditions, we subtracted the FFT grand-averaged spectra across conditions (congruent-incongruent) and calculated the baseline subtracted amplitude and the z-scores of the differential response. The topographical distribution revealed higher responses to congruent emotion expressions expressing over medial occipital, left occipito-temporal and fronto-central electrodes (positive amplitudes in figure 3B). 31 electrodes (i.e., 24.2188 % of the total set), mostly located over medial occipital, left occipito-temporal, left occipito-parietal and fronto-central scalp regions remained significant after FDR correction (PO7, PO9, PO11, I1,

POI1, POOz, O2, C2h, TP8h, FT8, FC6h, FC4, FC4h, FCC2h, FCC2, F8, FFC2, AFF2, AFF4h, Fp2, Afpz, Fz, F1, FCC1h, FCC1, FFC3h, FC3h, C1h, C3h, P7, P9; significant electrodes z-scores' range: 2.317142 - 5.795169, significant electrodes FDR-corrected p-values' range: 4.3682 x 10^{-7} - 0.0423; figures 3D, 3E). These results were further corroborated by the region-of-interest analysis. We observed a significant effect of condition on multisensory responses (congruent > incongruent) measured over occipito-parietal electrodes (vROI: t(18) = 2.3047, p = 0.0167 FDR corrected, Cohen's d = 0.5287; figure 3C) and over fronto-central electrodes (aROI: t(18) = 2.3326, p = 0.0167 FDR corrected, Cohen's d = 0.5351; figure 3C).

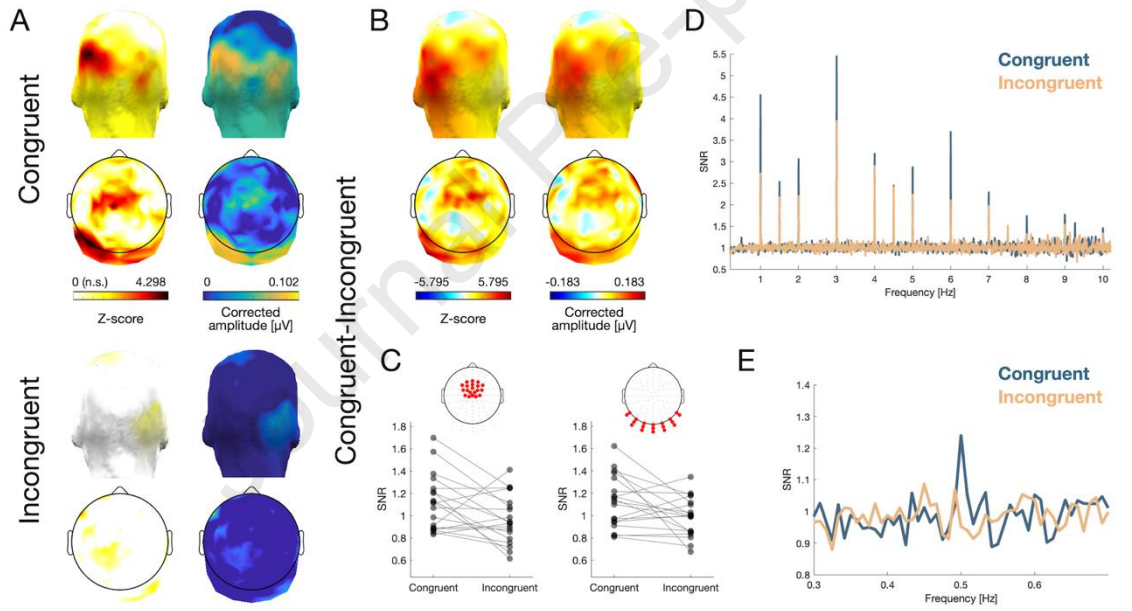


Figure 3. Intermodulation frequency responses to multisensory fearful expressions. A) Topographical responses at intermodulation frequencies components in the congruent (top) and incongruent (bottom) conditions. Maps on the left of each panel correspond to z-scores of the responses while maps on the right represent the baseline corrected amplitudes of the responses. Multisensory responses to facial and vocal affective expressions were observed in the congruent but not in the incongruent condition. B) Differential responses obtained subtracting EEG responses across conditions revealed higher responses when semantically congruent affective expressions were

presented peaking over 31 medial occipital, left occipito-temporal, left occipito-parietal and fronto-central electrodes. C) The region-of-interest analysis revealed higher responses to the congruent than to the incongruent condition in an auditory-defined ROI (left) and in a visually-defined ROI (right). Individual responses are displayed as dots connected by lines. D) SNR spectrum of the average of the 31 electrodes showing higher responses to congruent than incongruent affective expressions, with a zoom around 0.5 Hz ($\boldsymbol{f_{vis} - f_{aud}}$) in E (for visualization purposes).

4. Discussion

Consistent reports from behavioral studies suggest that integrating emotional expressions from the face and the voice into a fused and coherent percept is an automatic and mandatory process (Campanella & Belin, 2007; Collignon et al., 2008; Davies-Thompson et al., 2019; de Gelder et al., 2002; de Gelder & Vroomen, 2000; Dolan et al., 2001; Ethofer, Anders, et al., 2006; Ethofer, Pourtois, et al., 2006; Falagiarda & Collignon, 2019; Föcker & Röder, 2019; Klasen et al., 2012; Massaro & Egan, 1996; Pourtois et al., 2000, 2002; Watson et al., 2014). The core evolutionary role of multisensory integration to discriminate emotional expressions is supported by the fact that integrating facial and vocal components of socially relevant signals is anchored very early in life in humans (Grossmann et al., 2012; Hyde et al., 2016; Lewkowicz & Ghazanfar, 2009) and can be found in other nonhuman primates (Chandrasekaran et al., 2011; Ghazanfar & Chandrasekaran, 2012; Ghazanfar & Eliades, 2014). However, due to the intrinsic constraints of non-invasive neuroimaging techniques and the drawbacks of comparing unimodal and multimodal responses, revealing direct signatures of multisensory emotion integration in humans remains elusive (Giani et al., 2012; Gondan & Röder, 2006; Stanford & Stein, 2007; Stevenson et al., 2014).

To reveal a direct measure of multisensory integration of emotion expressions non-invasively in the human brain, we designed a multi-input frequency tagging experiment where dynamic fearful facial and vocal expressions were presented concurrently at different frequencies. If there are neural populations that process and integrate fearful information coming from the face and the voice, we expected to observe responses at intermodulation frequency components arising as a combination of the auditory and visual stimulation frequencies ($mf_{aud} \pm nf_{vis}$) (Baldauf & Desimone, 2014; Boremanse et al., 2013, 2014; Giani et al., 2012; Gordon et al., 2019; Morgan et al., 1996; Norcia et al., 2015; Regan et al., 1995). Considering that some brain regions respond similarly to emotionally congruent and incongruent bimodal stimuli (Klasen et al., 2012), to ensure that responses at intermodulation frequencies components would reflect integration of fearful expressions, we included a control condition in which fearful facial expression was paired with auditory disgust.

In both the congruent (visual fear / auditory fear) and incongruent (visual fear / auditory disgust) conditions we observed strong responses at the stimulation frequencies and related harmonics reflecting the processing of auditory and visual emotion expressions as well as more general low-level auditory and visual processes (figure 2). Interestingly, congruent voice emotion information was associated with an enhanced response to fearful facial expressions. That is, even though the visual stimulation was identical in the two conditions, there was a higher response in the congruent condition expressed in ~ 30 % of the electrodes and located over right/centro-frontal and left-medial occipital regions at the presentation frequency of the facial fearful expression. This crossmodal effect is in line with previous results reporting higher amplitudes when congruent facial and vocal affective information are presented together (Pourtois et al.,

2000). Moreover, as fearful facial expressions are more likely to be perceived as fearful if accompanied by a fearful voice (de Gelder & Vroomen, 2000; Dolan et al., 2001; Massaro & Egan, 1996), and since more intense facial affective expressions are associated with higher amplitude brain responses (Leleu et al., 2018), we could speculate that the higher visual responses observed in the congruent condition originate from the fearful facial expressions being perceived as more intense when accompanied with fearful voices than in the incongruent condition.

Crucially, our paradigm allowed us to isolate responses at intermodulation frequencies arising from the integration of affective information from the face and the voice. When presenting fearful facial and vocal expressions concurrently, we observed responses at intermodulation frequencies (IM) peaking at electrodes located over central and bilateral occipito-temporal scalp regions with a distinct topographical distribution from the unisensory responses. In contrast, there were no IM when fearful facial expressions were accompanied by disgust vocalizations, suggesting that the response observed in the congruent condition reflects integration of bimodal fearful information rather than mere selectivity to co-occurrent facial and vocal information. We further analyzed the difference between the two conditions by subtracting the two evoked responses. The differential response expressed over medial occipital, left occipito-temporal and fronto-central electrodes with a higher response to congruent than incongruent trials (figure 3B). These results were corroborated by a region-of-interest analysis: participants showed higher responses to the congruent than to the incongruent condition in auditory- and visual-defined ROIs (figure 3C). Since IM activity cannot be generated unless there are populations of neurons processing both stimulation frequency components (Gordon et al., 2019; Norcia et al., 2015; Regan et al., 1995), our study goes

beyond previous neuroimaging studies by pointing towards the existence of common neural populations that integrate fearful information from the face and the voice.

Responses at intermodulation frequencies have been repeatedly reported as evidence of integration processes occurring within senses (Alp & Ozkan, 2022; Baldauf & Desimone, 2014; Boremanse et al., 2013, 2014; Giani et al., 2012; Gordon et al., 2019; Goupil et al., 2023; Morgan et al., 1996; Norcia et al., 2015; Pitchaimuthu et al., 2021). Conversely, reports of intermodulation frequencies responses across senses are scarce, if non-existent. To our knowledge, the early results of Regan and collaborators about the integration of low-level auditory and visual information in two subjects have never been extended to a full study or replicated (Regan et al., 1995), and Giani and collaborators could identify intermodulation frequency responses within but not between sensory modalities (Giani et al., 2012). In a more recent study, Drijvers and colleagues reported intermodulation frequencies responses to uttered action verbs and semantically congruent gestures (Drijvers et al., 2021). However, the frequency-modulated inputs were the auditory and visual carriers of the stimulation rather than the stimuli themselves: in their experimental design, the luminance of a portion of the visual stimulation was tagged at 68 Hz and the speech was amplitude modulated at 61 Hz. In contrast, in our experiment, the tagging concerned the facial and voice emotions directly, therefore offering a direct link between the tagged responses at the cognitive processes at stake.

In our view, the scarcity of reports of multisensory intermodulation frequency responses might be attributable to the phenomenon investigated and the stimuli used. Because of its multisensory nature, bimodal emotion processing seems the ideal candidate to investigate the existence of neural populations processing and integrating audio-visual information. In fact, evidence from a wealth of behavioral and neuroimaging studies

suggest that integration of multimodal emotion information happens automatically, mandatorily and irrespective of attentional focus (Collignon et al., 2008; Klasen et al., 2012). In particular, among other emotion expressions, fear seems to be processed in priority, more rapidly and to have a privileged access to awareness, and previous studies suggest that the neural correlates of fear processing might differ from the processing of other affective information (Anderson et al., 2003; Dolan et al., 2001; Ethofer, Pourtois, et al., 2006; Falagiarda & Collignon, 2019; T. Stein et al., 2014).

It has been speculated that multisensory integration might be tuned to behaviorally relevant stimuli. Emotion discrimination is a dynamic process as affective expressions change from moment to moment (Hagan et al., 2009) and unfold over time (Falagiarda & Collignon, 2019; Jessen & Kotz, 2011). Therefore, the use of socially relevant and dynamic stimuli such as fearful expressions, as opposed to the use of lower-level multisensory stimuli which are hardly encountered concurrently in real life situations, might unveil sensory interactions that could otherwise go unnoticed when using more simplistic stimuli (Kayser, 2010). In this context, an interesting avenue for future studies would be to implement multi-input frequency tagging paradigms to investigate other aspects of multisensory person perception, such as the integration of audiovisual speech or identity. Moreover, the use of a multi-input frequency tagging method requires to present the facial and vocal emotion expressions at different presentation frequencies, disrupting a perfect synchrony of the visual and auditory information. If, on one side, stimuli onset asynchronies play a role in the perceptual binding of multisensory events and in the detection of multisensory responses, on the other side complex naturalistic multisensory stimuli show a larger temporal binding window (in other words, greater tolerance to stimuli asynchronies; Wallace & Stevenson, 2014). Lastly, the stimulation

frequencies were selected to capitalize on the principle of "inverse effectiveness" (B. E. Stein & Meredith, 1993) while taking into account methodological considerations (i.e., using close frequencies (Boremanse et al., 2013, 2014) and having responses not falling too low in the frequency spectrum to avoid substantial electrophysiological noise; Luck, 2014). Although the parameters and the stimuli implemented offered an optimal time window to investigate audio-visual integration of fearful expressions, further studies attempting to generalize our findings to other emotion expressions might have to fine-tune stimuli duration as different emotions need different time windows to be identified with equal accuracy (Collignon et al., 2008; Falagiarda & Collignon, 2019; Paulmann & Kotz, 2018).

To conclude, thanks to a multi-input frequency tagging approach we identified multisensory responses to bimodal fearful expressions. For the first time, we could isolate crossmodal intermodulation frequencies responses which were dependent on the congruency of simultaneously presented facial and vocal affective information, suggesting the existence of common neural populations in the human brain that mediate the processing of fearful expressions across the senses. Although further studies are needed to generalize our findings to other emotion expressions or to other multisensory integration processes in a broader sense, multi-input frequency tagging offers a promising tool to objectively and non-invasively identify multisensory integration within the same neural population in the human brain.

Transparency and openness promotion:

Original data is available here: https://data.mendeley.com/preview/54p8bpvw22?a=5b6e2ac3-1e82-4f17-b44a-7287fb910b98

If the manuscript will be accepted for publication, original code will be made available on a public repository.

No part of the study procedures and analyses was pre-registered prior to the research being conducted. We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study.

Acknowledgements:

This work was partially supported by the Belgian Excellence of Science program (EOS Project No. 30991544) attributed to OC & BR, the Flag-ERA HBP PINT-MULTI (R.8008.19) attributed to OC, a mandat d'impulsion scientifique (MIS-FNRS) attributed to OC and an ERC Adg (HUMANFACE 101055175) to BR. FMB, ST and RPC are research fellows and OC a senior research associate (maître de recherche) at the National Fund for Scientific Research of Belgium (FRS-FNRS). We would like to thank Professor Arnaud Leleu, Dr. Diane Rekow and Anna Kiseleva for their insightful comments on the data. We are grateful to Professor Sylvie Nozaradan, Professor André Mouraux, Professor Davide Bottari, Professor Mathieu Bourguignon and Professor Jolijn Vanderauwera for their precious feedback on an early version of the manuscript. Finally, we would like to extend our gratitude to Marie-Ange Azoury for her help in the data acquisition.

Author contributions:

FMB: Conceptualization; Methodology; Software; Data curation; Formal analysis; Investigation; Validation; Visualization; Writing – original draft; Writing – review & editing; Project administration.

ST: Software; Formal analysis; Investigation; Writing – review & editing.

RPC: Investigation; Writing – review & editing.

BR: Conceptualization; Methodology; Validation; Supervision; Writing – review & editing.

OC: Conceptualization; Methodology; Funding acquisition; Project administration; Resources; Supervision; Validation; Writing - review & editing.

References

Alp, N., & Ozkan, H. (2022). Neural correlates of integration processes during dynamic face perception. *Scientific Reports*, *12*(1), 118. https://doi.org/10.1038/s41598-021-02808-9

Anderson, A. K., Christoff, K., Panitz, D., De Rosa, E., & Gabrieli, J. D. E. (2003). Neural Correlates of the Automatic Processing of Threat Facial Signals. *The Journal of Neuroscience*, *23*(13), 5627–5633. https://doi.org/10.1523/JNEUROSCI.23-13-05627.2003

Baldauf, D., & Desimone, R. (2014). Neural Mechanisms of Object-Based Attention. *Science*, *344*(6182), 424–427. https://doi.org/10.1126/science.1247003

Bänziger, T., Grandjean, D., & Scherer, K. R. (2009). Emotion recognition from expressions in face, voice, and body: The Multimodal Emotion Recognition Test (MERT). *Emotion*, *9*(5), 691–704. https://doi.org/10.1037/a0017088

Barbero, F. M., Calce, R. P., Talwar, S., Rossion, B., & Collignon, O. (2021). Fast Periodic Auditory Stimulation Reveals a Robust Categorical Response to Voices in the Human Brain. *Eneuro*, *8*(3), ENEURO.0471-20.2021. https://doi.org/10.1523/ENEURO.0471-20.2021

Barraclough*, N. E., Xiao*, D., Baker, C. I., Oram, M. W., & Perrett, D. I. (2005). Integration of Visual and Auditory Information by Superior Temporal Sulcus

Neurons Responsive to the Sight of Actions. *Journal of Cognitive Neuroscience*, *17*(3), 377–391. https://doi.org/10.1162/0898929053279586

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal Affective Voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods*, *40*(2), 531–539. https://doi.org/10.3758/BRM.40.2.531

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Boremanse, A., Norcia, A. M., & Rossion, B. (2013). An objective signature for visual binding of face parts in the human brain. *Journal of Vision*, *13*(11), 6–6. https://doi.org/10.1167/13.11.6

Boremanse, A., Norcia, A. M., & Rossion, B. (2014). Dissociation of part-based and integrated neural responses to faces by means of electroencephalographic frequency tagging. *European Journal of Neuroscience*, *40*(6), 2987–2997. https://doi.org/10.1111/ejn.12663

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Calce, R. P., Rekow, D., Barbero, F. M., Kiseleva, A., Talwar, S., Leleu, A., & Collignon, O. (2024). Voice categorization in the four-month-old human brain. *Current Biology*, *34*(1), 46-55.e4. https://doi.org/10.1016/j.cub.2023.11.042

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*(12), 535–543. https://doi.org/10.1016/j.tics.2007.10.001

Chandrasekaran, C., Lemus, L., Trubanova, A., & Gondan, M. (2011). Monkeys and Humans Share a Common Computation for Face/Voice Integration. *PLoS Computational Biology*, *7*(9), 20.

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research*, *1242*, 126–135. https://doi.org/10.1016/j.brainres.2008.04.023

Darwin, C. (1872). *The expression of the emotions in man and animals.* John Murray. https://doi.org/10.1037/10001-000

Davies-Thompson, J., Elli, G. V., Rezk, M., Benetti, S., van Ackeren, M., & Collignon, O. (2019). Hierarchical Brain Network for Face and Voice Integration of Emotion Expression. *Cerebral Cortex*, *29*(9), 3590–3605. https://doi.org/10.1093/cercor/bhy240

de Gelder, B., Böcker, K. B. E., Tuomainen, J., Hensen, M., & Vroomen, J. (1999). The combined perception of emotion from voice and face: Early interaction revealed by human electric brain responses. *Neuroscience Letters*, *260*(2), 133–136. https://doi.org/10.1016/S0304-3940(98)00963-X

de Gelder, B., Pourtois, G., & Weiskrantz, L. (2002). Fear recognition in the voice is modulated by unconsciously recognized facial expressions but not by unconsciously recognized affective pictures. *Proceedings of the National Academy of Sciences*, *99*(6), 4121–4126. https://doi.org/10.1073/pnas.062018499

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion*, *14*(3), 289–311. https://doi.org/10.1080/026999300378824

Dolan, R. J., Morris, J. S., & de Gelder, B. (2001). Crossmodal binding of fear in voice and face. *Proceedings of the National Academy of Sciences*, *98*(17), 10006–10010. https://doi.org/10.1073/pnas.171288598

Drijvers, L., Jensen, O., & Spaak, E. (2021). Rapid invisible frequency tagging reveals nonlinear integration of auditory and visual information. *Human Brain Mapping*, *42*(4), 1138–1152. https://doi.org/10.1002/hbm.25282

Dzhelyova, M., Jacques, C., & Rossion, B. (2016). At a Single Glance: Fast Periodic Visual Stimulation Uncovers the Spatio-Temporal Dynamics of Brief Facial Expression Changes in the Human Brain. *Cerebral Cortex*, cercor;bhw223v1. https://doi.org/10.1093/cercor/bhw223

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, *1*(1), 56–75. https://doi.org/10.1007/BF01115465

Ethofer, T., Anders, S., Erb, M., Droll, C., Royen, L., Saur, R., Reiterer, S., Grodd, W., & Wildgruber, D. (2006). Impact of voice on emotional judgment of faces: An event-related fMRI study. *Human Brain Mapping*, *27*(9), 707–714. https://doi.org/10.1002/hbm.20212

Ethofer, T., Pourtois, G., & Wildgruber, D. (2006). Investigating audiovisual integration of emotional signals in the human brain. In *Progress in Brain Research* (Vol. 156, pp. 345–361). Elsevier. https://doi.org/10.1016/S0079-6123(06)56019-4

Falagiarda, F., & Collignon, O. (2019). Time-resolved discrimination of audio-visual emotion expressions. *Cortex*, *119*, 184–194. https://doi.org/10.1016/j.cortex.2019.04.017

Föcker, J., & Röder, B. (2019). Event-Related Potentials Reveal Evidence for Late Integration of Emotional Prosody and Facial Expression in Dynamic Stimuli: An ERP Study. *Multisensory Research*, *32*(6), 473–497. https://doi.org/10.1163/22134808-20191332

Ghazanfar, A. A., & Chandrasekaran, C. (2012). Multisensory Vocal Communication and Its Neural Bases in Nonhuman Primates. In B. E. Stein (Ed.), *The New Handbook of Multisensory Processing* (pp. 407–420). The MIT Press. https://doi.org/10.7551/mitpress/8466.003.0036

Ghazanfar, A. A., & Eliades, S. J. (2014). The neurobiology of primate vocal communication. *Current Opinion in Neurobiology*, *28*, 128–135. https://doi.org/10.1016/j.conb.2014.06.015

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory Integration of Dynamic Faces and Voices in Rhesus Monkey Auditory Cortex. *The Journal of Neuroscience*, *25*(20), 5004–5012. https://doi.org/10.1523/JNEUROSCI.0799-05.2005

Ghazanfar, A. A., & Schroeder, C. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, *10*(6), 278–285. https://doi.org/10.1016/j.tics.2006.04.008

Giani, A. S., Ortiz, E., Belardinelli, P., Kleiner, M., Preissl, H., & Noppeney, U. (2012). Steady-state responses in MEG demonstrate information integration within but not across the auditory and visual senses. *NeuroImage*, *60*(2), 1478–1489. https://doi.org/10.1016/j.neuroimage.2012.01.114

Gondan, M., & Röder, B. (2006). A new method for detecting interactions between the senses in event-related potentials. *Brain Research*, *1073–1074*, 389–397. https://doi.org/10.1016/j.brainres.2005.12.050

Gordon, N., Hohwy, J., Davidson, M. J., van Boxtel, J. J. A., & Tsuchiya, N. (2019). From intermodulation components to visual perception and cognition-a review. *NeuroImage*, *199*, 480–494. https://doi.org/10.1016/j.neuroimage.2019.06.008

Goupil, N., Hochmann, J.-R., & Papeo, L. (2023). Intermodulation responses show integration of interacting bodies in a new whole. *Cortex*, *165*, 129–140. https://doi.org/10.1016/j.cortex.2023.04.013

Grossmann, T., Missana, M., Friederici, A. D., & Ghazanfar, A. A. (2012). Neural correlates of perceptual narrowing in cross-species face-voice matching: ERPs and face-voice matching. *Developmental Science*, *15*(6), 830–839. https://doi.org/10.1111/j.1467-7687.2012.01179.x

Hagan, C. C., Woods, W., Johnson, S., Calder, A. J., Green, G. G. R., & Young, A. W. (2009). MEG demonstrates a supra-additive response to facial and vocal emotion in the right superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *106*(47), 20010–20015. https://doi.org/10.1073/pnas.0905792106

Ho, H. T., Schröger, E., & Kotz, S. A. (2015). Selective Attention Modulates Early Human Evoked Potentials during Emotional Face–Voice Processing. *Journal of Cognitive Neuroscience*, *27*(4), 798–818. https://doi.org/10.1162/jocn_a_00734

Hyde, D. C., Flom, R., & Porter, C. L. (2016). Behavioral and Neural Foundations of Multisensory Face-Voice Perception in Infancy. *Developmental Neuropsychology*, *41*(5–8), 273–292. https://doi.org/10.1080/87565641.2016.1255744

Jessen, S., & Kotz, S. A. (2011). The temporal dynamics of processing emotions from vocal, facial, and bodily expressions. *NeuroImage*, *58*(2), 665–674. https://doi.org/10.1016/j.neuroimage.2011.06.035

Kayser, C. (2010). The Multisensory Nature of Unisensory Cortices: A Puzzle Continued. *Neuron*, *67*(2), 178–180. https://doi.org/10.1016/j.neuron.2010.07.012

Khandhadia, A. P., Murphy, A. P., Romanski, L. M., Bizley, J. K., & Leopold, D. A. (2021). Audiovisual integration in macaque face patch neurons. *Current Biology*, S0960982221001676. https://doi.org/10.1016/j.cub.2021.01.102

Klasen, M., Chen, Y.-H., & Mathiak, K. (2012). Multisensory emotions: Perception, combination and underlying neural processes. *Reviews in the Neurosciences*, *23*(4). https://doi.org/10.1515/revneuro-2012-0040

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, *36*(14), 1–16.

Klinge, C., Röder, B., & Büchel, C. (2010). Increased amygdala activation to emotional auditory stimuli in the blind. *Brain*, *133*(6), 1729–1736. https://doi.org/10.1093/brain/awq102

Kokinous, J., Kotz, S. A., Tavano, A., & Schröger, E. (2015). The role of emotion in dynamic audiovisual integration of faces and voices. *Social Cognitive and Affective Neuroscience*, *10*(5), 713–720. https://doi.org/10.1093/scan/nsu105

Leleu, A., Dzhelyova, M., Rossion, B., Brochard, R., Durand, K., Schaal, B., & Baudouin, J.-Y. (2018). Tuning functions for automatic detection of brief changes of facial expression in the human brain. *NeuroImage*, *179*, 235–251. https://doi.org/10.1016/j.neuroimage.2018.06.048

Lewkowicz, D. J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. *Trends in Cognitive Sciences*, *13*(11), 470–478. https://doi.org/10.1016/j.tics.2009.08.004

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*(7197), 869–878. https://doi.org/10.1038/nature06976

Luck, S. J. (2014). *An introduction to the event-related potential technique* (Second edition). The MIT Press.

Massaro, D. W., & Egan, P. B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review*, *3*(2), 215–221. https://doi.org/10.3758/BF03212421

Matt, S., Dzhelyova, M., Maillard, L., Lighezzolo-Alnot, J., Rossion, B., & Caharel, S. (2021). The rapid and automatic categorization of facial expression changes in highly variable natural images. *Cortex*, *144*, 168–184. https://doi.org/10.1016/j.cortex.2021.08.005

Morgan, S. T., Hansen, J. C., & Hillyard, S. A. (1996). Selective attention to stimulus location modulates the steady-state visual evoked potential. *Proceedings of the National Academy of Sciences*, *93*(10), 4770–4774. https://doi.org/10.1073/pnas.93.10.4770

Norcia, A. M., Appelbaum, L. G., Ales, J. M., Cottereau, B. R., & Rossion, B. (2015). The steady-state visual evoked potential in vision research: A review. *Journal of Vision*, *15*(6), 4. https://doi.org/10.1167/15.6.4

Paulmann, S., & Kotz, S. A. (2018). The Electrophysiology and Time Course of Processing Vocal Emotion Expressions. In S. Frühholz & P. Belin (Eds.), *The*

Oxford Handbook of Voice Perception (pp. 458–472). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198743187.013.20

Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal Representations of Perceived Emotions in the Human Brain. *Journal of Neuroscience*, *30*(30), 10127–10134. https://doi.org/10.1523/JNEUROSCI.2161-10.2010

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Perrodin, C., Kayser, C., Logothetis, N. K., & Petkov, C. I. (2014). Auditory and Visual Modulation of Temporal Lobe Neurons in Voice-Sensitive and Association Cortices. *Journal of Neuroscience*, *34*(7), 2524–2537. https://doi.org/10.1523/JNEUROSCI.2805-13.2014

Perrodin, C., Kayser, C., Logothetis, N. K., & Petkov, C. I. (2015). Natural asynchronies in audiovisual communication signals regulate neuronal multisensory interactions in voice-sensitive cortex. *Proceedings of the National Academy of Sciences*, *112*(1), 273–278. https://doi.org/10.1073/pnas.1412817112

Pitchaimuthu, K., Dormal, G., Sourav, S., Shareef, I., Rajendran, S. S., Ossandón, J. P., Kekunnaya, R., & Röder, B. (2021). Steady state evoked potentials indicate changes in nonlinear neural mechanisms of vision in sight recovery individuals. *Cortex*, *144*, 15–28. https://doi.org/10.1016/j.cortex.2021.08.001

Pourtois, G., de Gelder, B., Vroomen, J., Rossion, B., & Crommelinck, M. (2000). The time-course of intermodal binding between seeing and hearing affective information: *NeuroReport*, *11*(6), 1329–1333. https://doi.org/10.1097/00001756-200004270-00036

Pourtois, G., Debatisse, D., Despland, P.-A., & de Gelder, B. (2002). Facial expressions modulate the time course of long latency auditory brain potentials. *Cognitive Brain Research*, *14*(1), 99–105. https://doi.org/10.1016/S0926-6410(02)00064-2

Pourtois, G., Degelder, B., Bol, A., & Crommelinck, M. (2005). Perception of Facial Expressions and Voices and of their Combination in the Human Brain. *Cortex*, *41*(1), 49–59. https://doi.org/10.1016/S0010-9452(08)70177-1

Regan, M. P., Regan, D., & He, P. (1995). An audio-visual convergence area in the human brain. *Experimental Brain Research*, *106*(3). https://doi.org/10.1007/BF00231071

Retter, T. L., & Rossion, B. (2016). Uncovering the neural magnitude and spatio-temporal dynamics of natural image categorization in a fast visual stream. *Neuropsychologia*, *91*, 9–28. https://doi.org/10.1016/j.neuropsychologia.2016.07.028

Retter, T. L., Rossion, B., & Schiltz, C. (2021). Harmonic Amplitude Summation for Frequency-tagging Analysis. *Journal of Cognitive Neuroscience*, 1–22. https://doi.org/10.1162/jocn_a_01763

Romanski, L. M. (2007). Representation and Integration of Auditory and Visual Stimuli in the Primate Ventral Lateral Prefrontal Cortex. *Cerebral Cortex*, *17*(suppl 1), i61–i69. https://doi.org/10.1093/cercor/bhm099

Romanski, L. M. (2012). Integration of faces and vocalizations in ventral prefrontal cortex: Implications for the evolution of audiovisual speech. *Proceedings of the National Academy of Sciences*, *109*(supplement_1), 10717–10724. https://doi.org/10.1073/pnas.1204335109

Rstudio, T. (2020). RStudio: Integrated Development for R. In *Rstudio Team, PBC, Boston, MA URL http://www.rstudio.com/*. https://doi.org/10.1145/3132847.3132886

Simon, D., Craig, K. D., Miltner, W. H. R., & Rainville, P. (2006). Brain responses to dynamic facial expressions of pain. *Pain*, *126*(1), 309–318. https://doi.org/10.1016/j.pain.2006.08.033

Stanford, T. R., & Stein, B. E. (2007). Superadditivity in multisensory integration: Putting the computation in context. *NeuroReport*, *18*(8), 787–792. https://doi.org/10.1097/WNR.0b013e3280c1e315

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses*. MIT Press.

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, *9*(4), 255–266. https://doi.org/10.1038/nrn2331

Stein, T., Seymour, K., Hebart, M. N., & Sterzer, P. (2014). Rapid Fear Detection Relies on High Spatial Frequencies. *Psychological Science*, *25*(2), 566–574. https://doi.org/10.1177/0956797613512509

Stevenson, R. A., Ghose, D., Fister, J. K., Sarko, D. K., Altieri, N. A., Nidiffer, A. R., Kurela, L. R., Siemann, J. K., James, T. W., & Wallace, M. T. (2014). Identifying and Quantifying Multisensory Integration: A Tutorial Review. *Brain Topography*, *27*(6), 707–730. https://doi.org/10.1007/s10548-014-0365-7

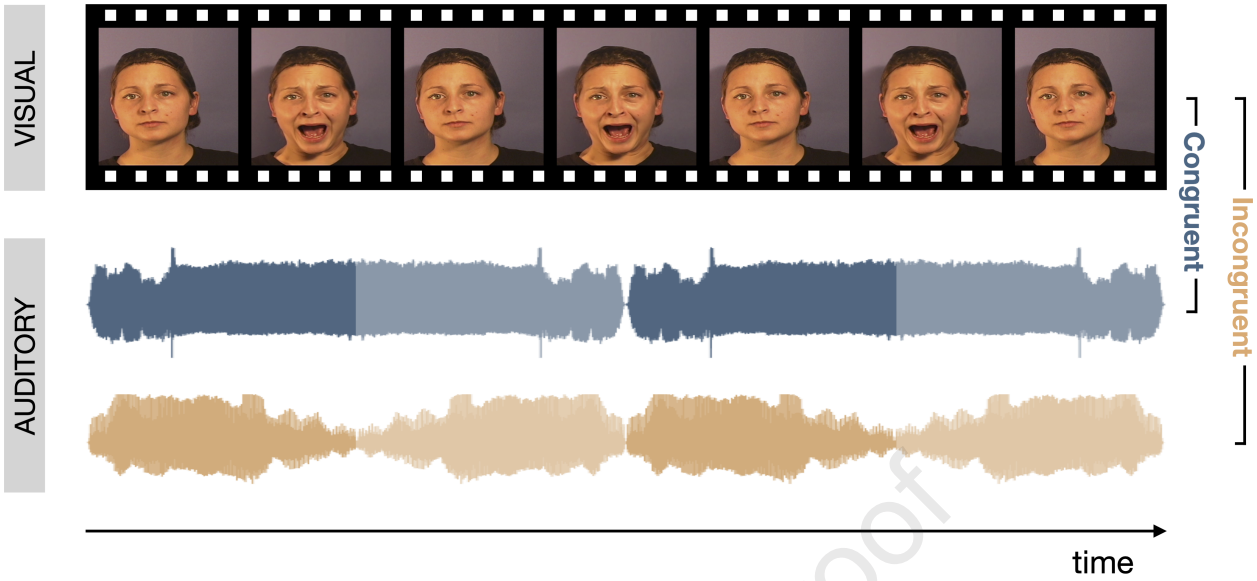
Sugihara, T., Diltz, M. D., Averbeck, B. B., & Romanski, L. M. (2006). Integration of Auditory and Visual Communication Information in the Primate Ventrolateral Prefrontal Cortex. *The Journal of Neuroscience*, *26*(43), 11138–11147. https://doi.org/10.1523/JNEUROSCI.3550-06.2006

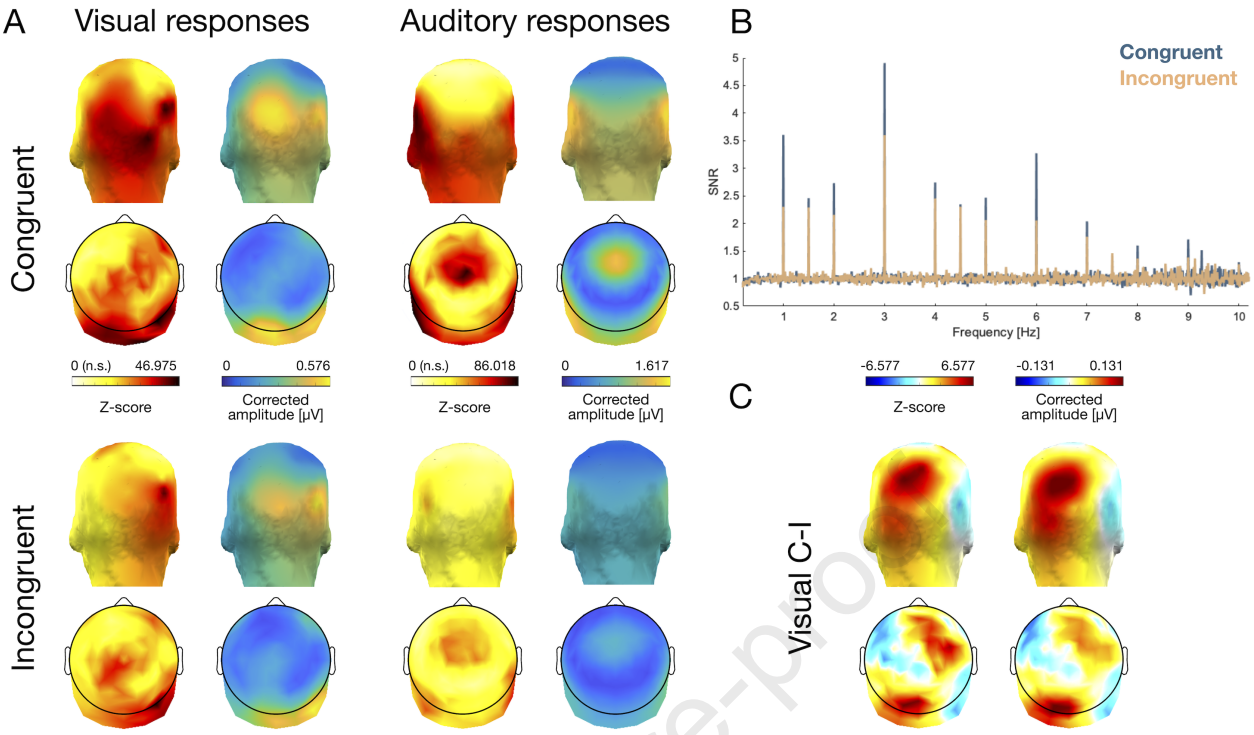
Talwar, S., Barbero, F. M., Calce, R. P., & Collignon, O. (2023). Automatic Brain Categorization of Discrete Auditory Emotion Expressions. *Brain Topography*, *36*(6), 854–869. https://doi.org/10.1007/s10548-023-00983-8

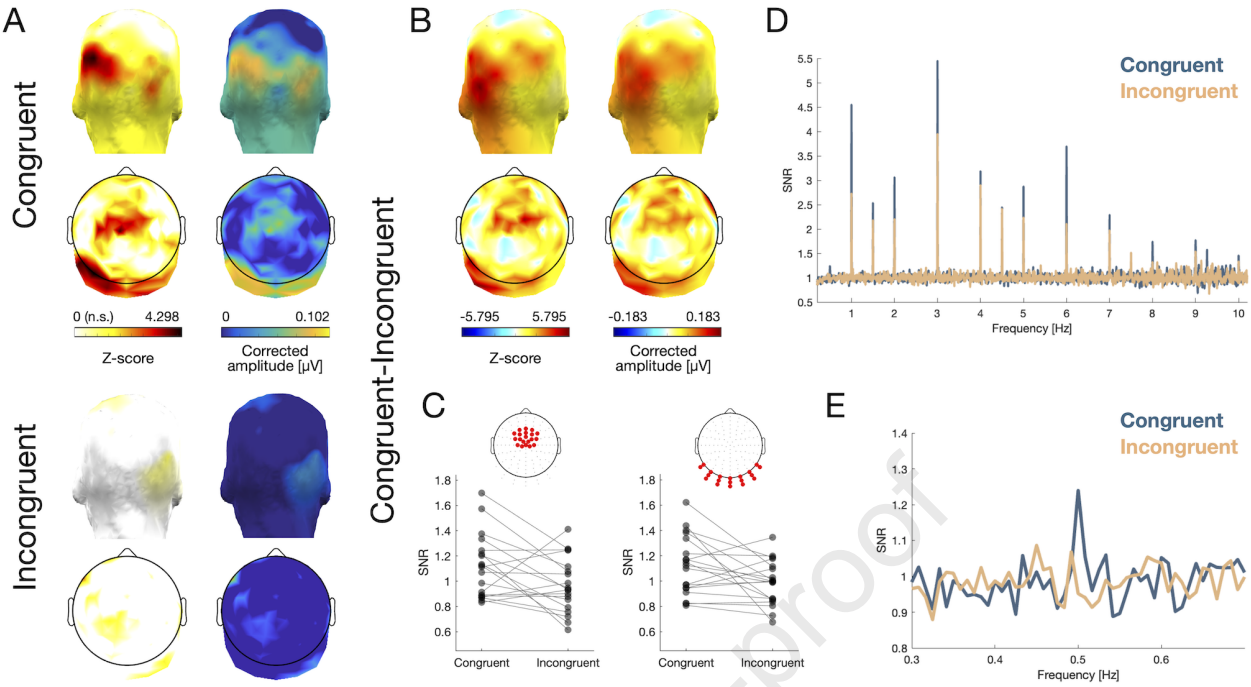
Volfart, A., Rice, G. E., Lambon Ralph, M. A., & Rossion, B. (2021). Implicit, automatic semantic word categorisation in the left occipito-temporal cortex as revealed by fast periodic visual stimulation. *NeuroImage*, *238*, 118228. https://doi.org/10.1016/j.neuroimage.2021.118228

Wallace, M. T., & Stevenson, R. A. (2014). The construct of the multisensory temporal binding window and its dysregulation in developmental disabilities. *Neuropsychologia*, *64*, 105–123. https://doi.org/10.1016/j.neuropsychologia.2014.08.005

Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., & Belin, P. (2014). Crossmodal Adaptation in Right Posterior Superior Temporal Sulcus during Face-Voice Emotional Integration. *Journal of Neuroscience*, *34*(20), 6813–6821. https://doi.org/10.1523/JNEUROSCI.4478-13.2014







Highlights

- Common neural populations integrate facial and vocal fearful expressions
- EEG frequency-tagging to provide a measure of audio-visual integration of fear
- Intermodulation frequencies as a non-invasive measure of integration in humans
- Intermodulation frequencies absent with mismatched audio-visual emotion expressions

Scientific transparency statement

DATA: All raw and processed data supporting this research are publicly available: https://osf.io/c6q4d/

CODE: All analysis code supporting this research is publicly available: https://osf.io/c6q4d/

MATERIALS: All study materials supporting this research are publicly available: https://osf.io/c6q4d/

DESIGN: This article reports, for all studies, how the author(s) determined all sample sizes, all data exclusions, all data inclusion and exclusion criteria, and whether inclusion and exclusion criteria were established prior to data analysis.

PRE-REGISTRATION: No part of the study procedures was pre-registered in a time-stamped, institutional registry prior to the research being conducted. No part of the analysis plans was pre-registered in a time-stamped, institutional registry prior to the research being conducted.

For full details, see the *Scientific Transparency Report* in the supplementary data to the online version of this article.